NIST Machine Translation 2004 Evaluation Summary of Results

NIST MT04 Evaluation Workshop

Mark Przybocki

June 22nd – 23rd, 2004

Hilton Alexandria Mark Center

Alexandria, Virginia

Outline

- Evaluation Conditions
- Evaluation Data
  - Source
  - Reference
- Performance Measurement
- Results
  - 2004 MT Evaluation
  - Significance Test
  - Evaluation History
- Summary Statements

Evaluation Conditions

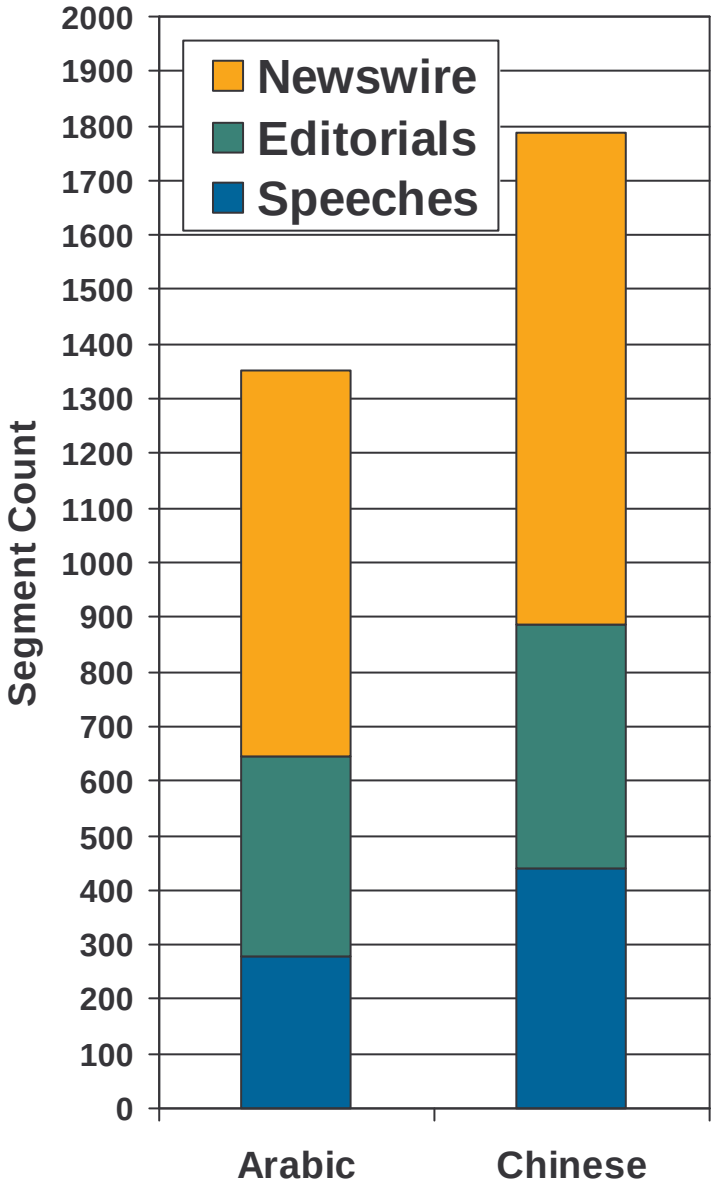
Source Language	Evaluation Condition	Restrictions on Resources for System Development
Arabic	Open	Cut-off date of Jan. 1, 2004
	Large	LDC provided resources
Chinese	Open	Cut-off date of Jan. 1, 2004
	Large	LDC provided resources
	Small	10K CMU dictionary 100K-word UPENN Chinese Treebank

The Jan. 1, 2004 cut-off date is established to give the LDC ample time to collect and transcribe evaluation data that won't accidentally be used for system development.

Evaluation Source Data

Language	Data Genre	Doc Count	Dates
Arabic *(UTF-8)*	Newswire	100	Jan 2004
	Editorials	50	Jan-Mar, 2004
	Speeches	50	2002-2004
Chinese *(GB)*	Newswire	100	Jan 2004
	Editorials	50	Nov 03 – Mar 04
	Speeches	50	2002 – 2004

- The collection of newswire is similar to past MT test sets
  - in size and date span (1 month)
- First year of different genre data
- More segments in the Chinese data



Evaluation Reference Data

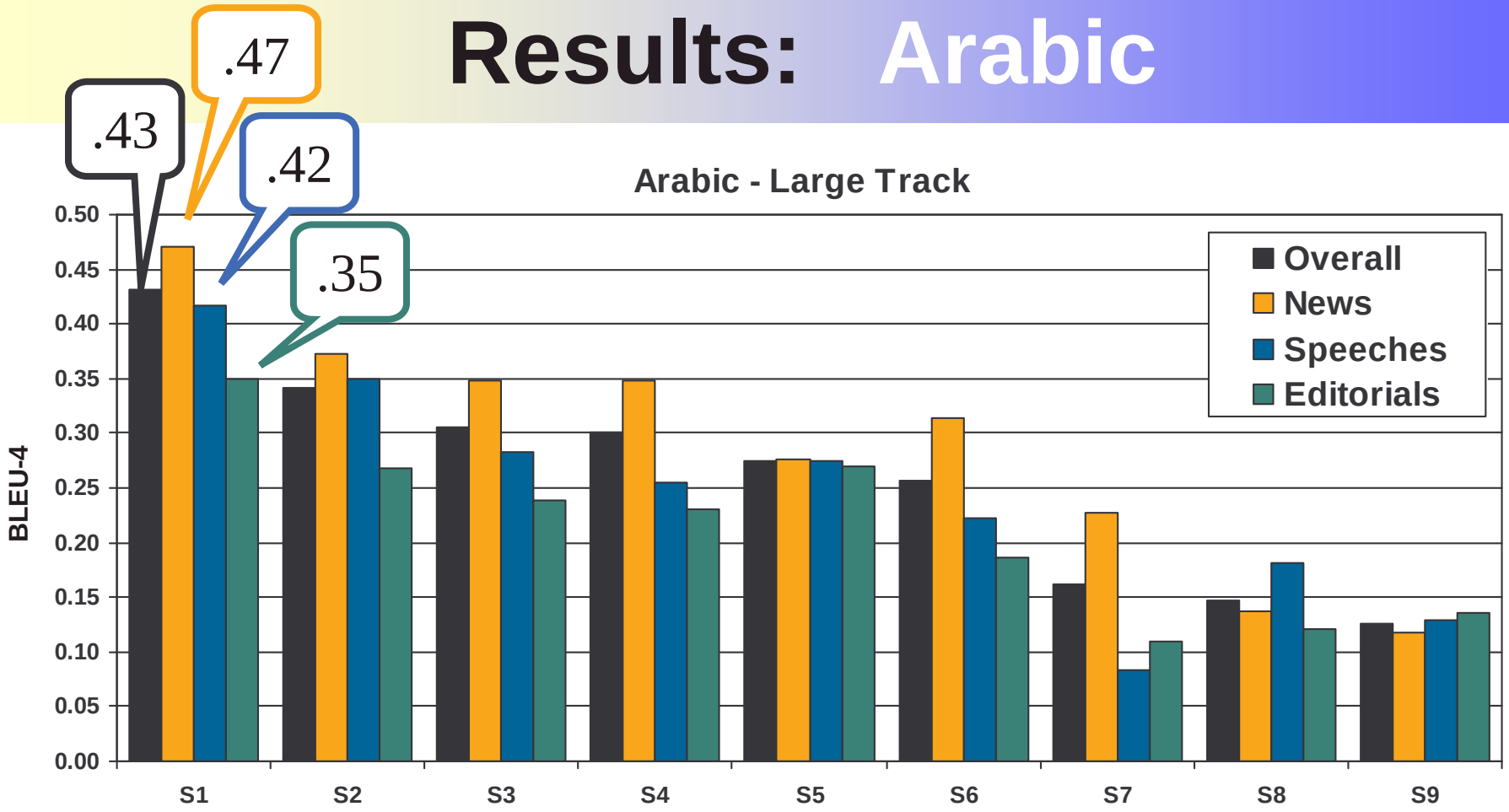
- Arabic
  - Four sets of human translations
    - 2 of the 4 came from the same translation agencies as last year
  - Average segment length of newswire translations comparable to 2003 (at ~28 words per segment)
- Chinese
  - Four sets of human translations
    - 3 of the 4 came from the same translation agencies as last year
  - Average segment length of newswire translations comparable to 2003 (at ~28 words per segment)

Performance Measurement

- BLEU Metric as proposed by IBM[1]
  - Expressed in terms of an un-weighted geometric average of n-gram precisions, for n=1,2,3,4
  - Includes a penalty for translations whose length differs significantly from that of the reference translations
- Case Sensitive Scoring

[1] Papineni, Roukos, Ward, Zhu (2001). “Bleu: a Method for Automatic Evaluation of Machine Translation”.

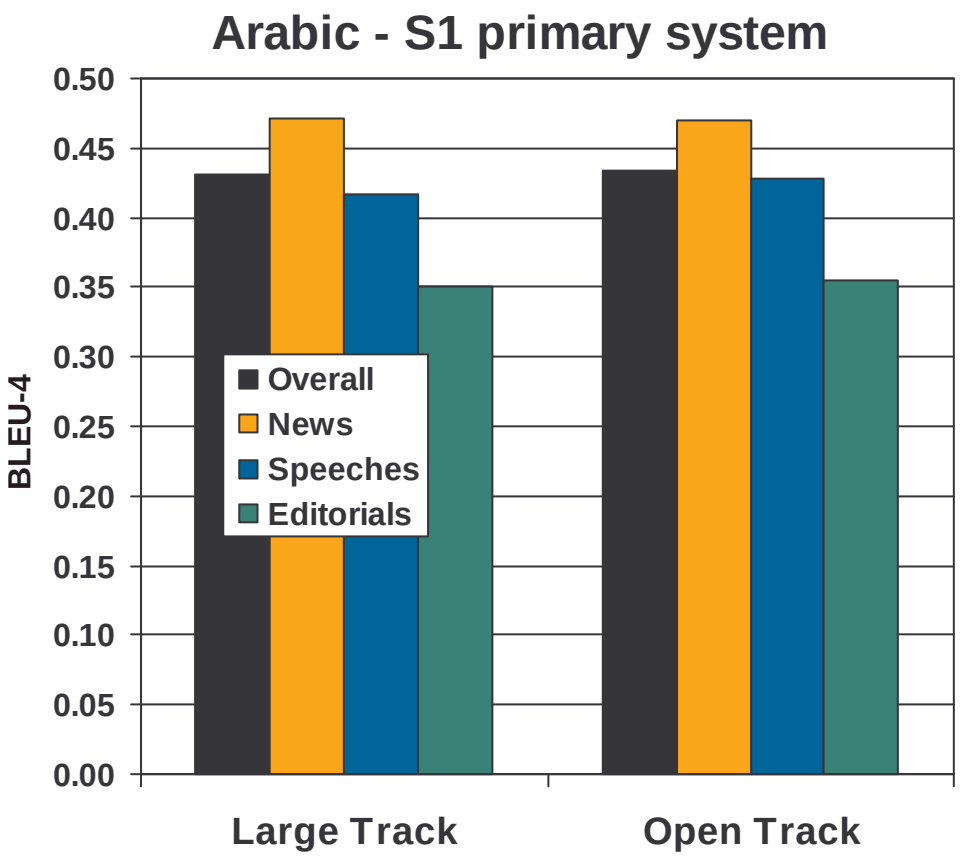
Results: Arabic



- Primary Systems ordered by overall MT04 test set score

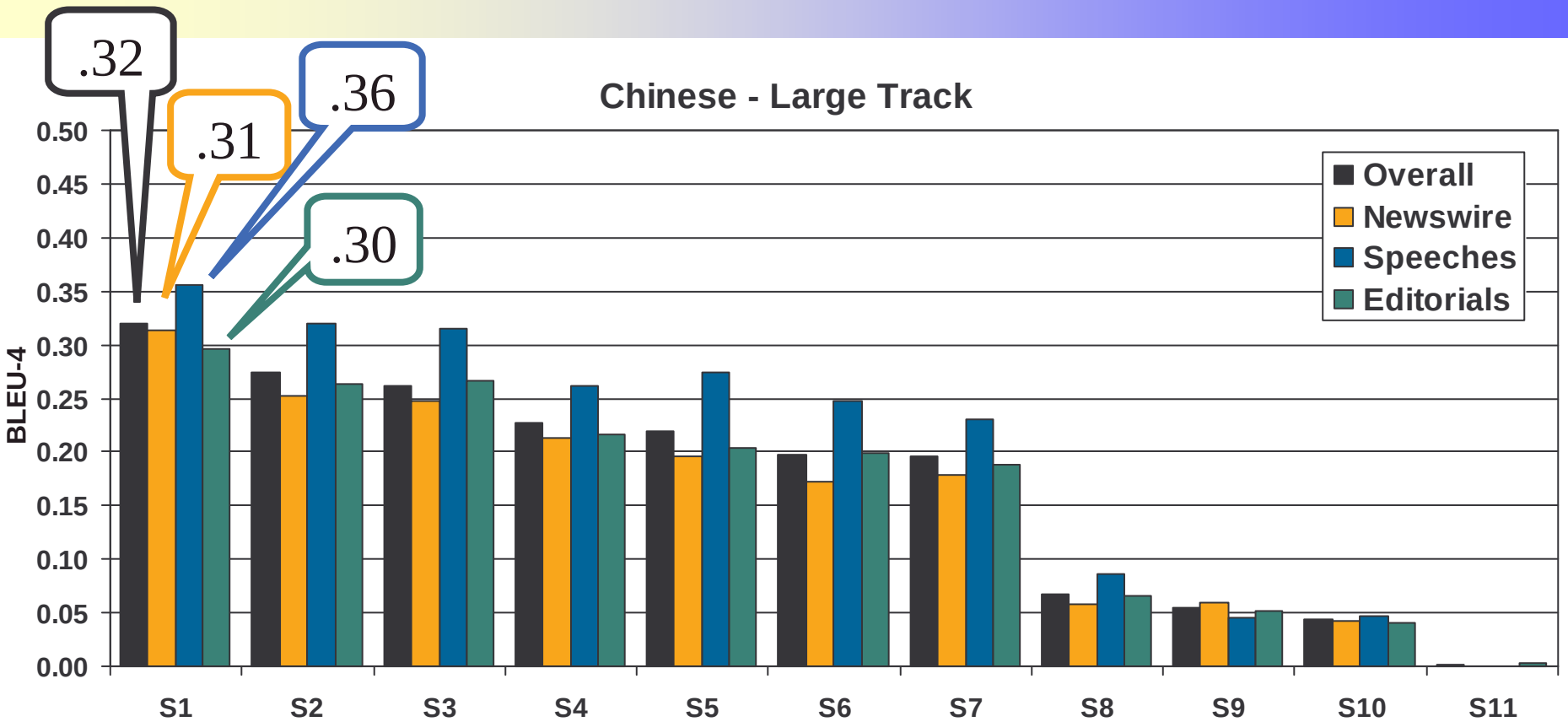
- Ranking the genre difficulty
  - Newswire (least difficult)
  - Combined test set
  - Speeches
  - Editorials (most difficult)

Results: Arabic (cont'd)



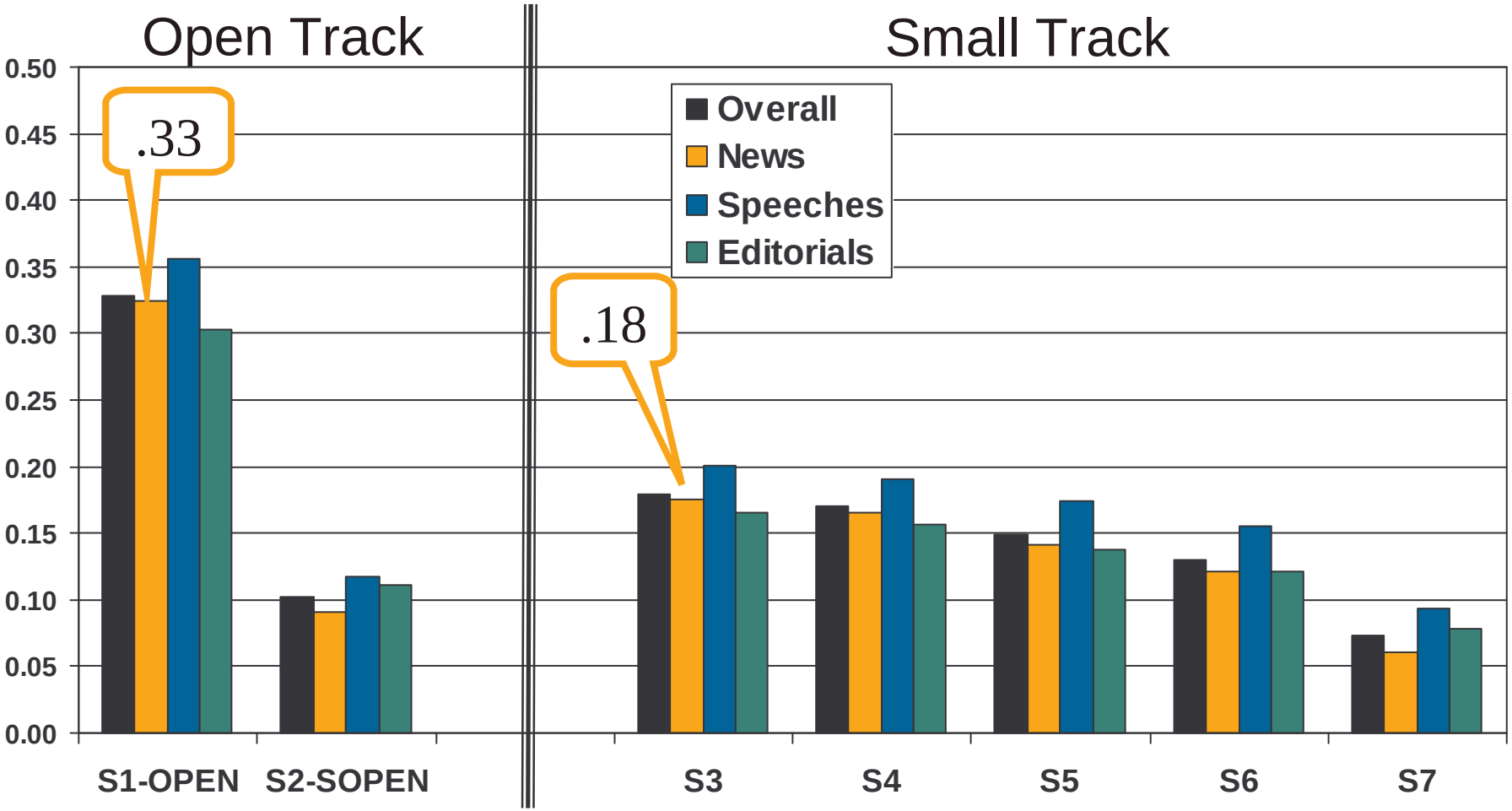
- S1 primary system for the large and open data track
- *Slight improvement* in Open track (.435) over Large track (.431) obtained for S1

Results: Chinese



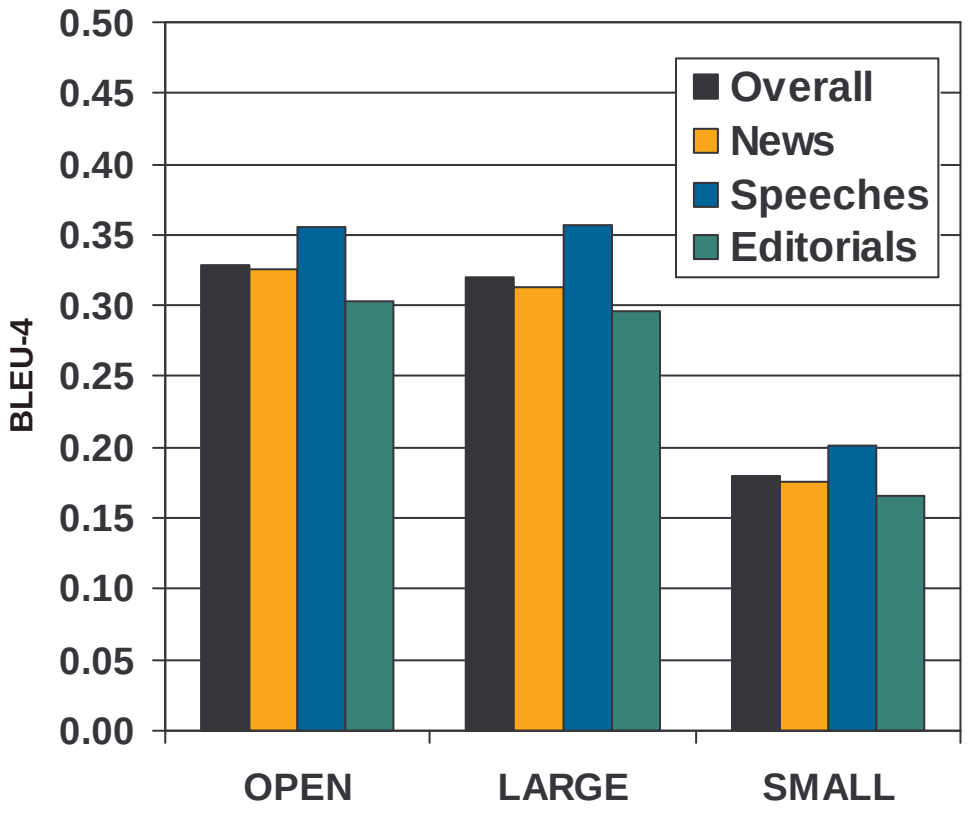
- Primary Systems ordered by overall MT04 test set score
- Ranking the genre difficulty
  - Speeches (least difficult)
  - Combined test set
  - Editorials
  - Newswire (most difficult)

Results: Chinese (cont'd)



- Primary Systems ordered by overall MT04 test set score
- Genre difficult consistent trend for most systems

Results: Chinese (cont'd)



- On site's primary system for each evaluation condition
- *Slight improvement* in Open track (.325) over Large track (.314) obtained on newswire data for one system
- Genre difficulty seems to remain constant across different evaluation conditions

Significance Test: Sign Test

- Compare BLEU scores on a per document basis
  - 200 trials (*entire test set being considered*)
  - For each document, increment ***site1-count*** if site1 has a higher BLEU score, else if site2 is higher, increment ***site2-count***
- Apply sign test for paired comparisons of BLEU scores
- The Null Hypothesis:
  - Site1 and site2 are **equally likely** to receive a higher BLEU score on a given document
- Alternate Hypothesis:
  - Site1 and site2 are **NOT equally likely** to receive a higher BLEU score on a given document

Sign Test Results: Arabic
Primary Systems for the LARGE data track

System Y	X1	X2	X3	X4	X5	X6	X7	X8	X9
Y1 (3)			<0.1%	<0.1%	<0.1%	>99.9%	<0.1%	>99.9%	>99.9%
Y2 (3)			0.6%	<0.1%	<0.1%	>99.9%	<0.1%	>99.9%	>99.9%
Y3 (5)				<0.1%		>99.9%	<0.1%	>99.9%	>99.9%
Y4 (8)					>99.9%	>99.9%	>99.9%	>99.9%	>99.9%
Y5 (5)						>99.9%	<0.1%	>99.9%	>99.9%
Y6 (0)							<0.1%		0.1%
Y7 (7)								>99.9%	>99.9%
Y8 (0)									0.2%
Y9 (2)									

(5) – The number of comparisons (out of 8) where Y5 was determined to be the better system

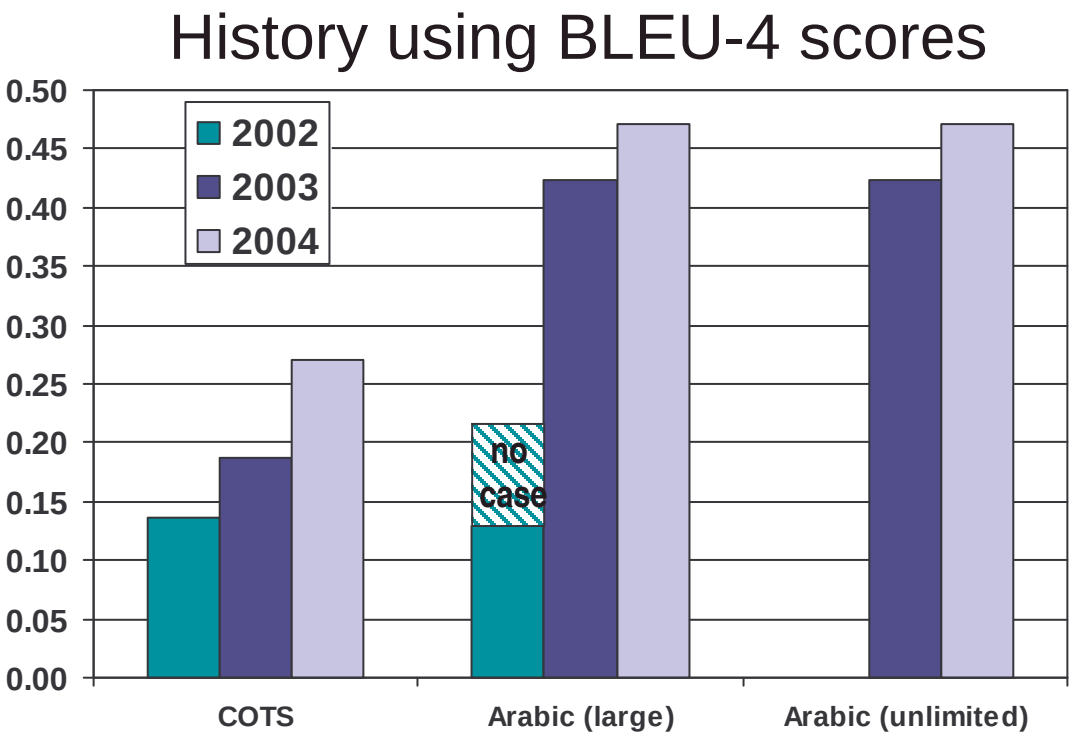
- Answers the question:
  What is the confidence that system Y is better than system X?
- If no entry (yellow cells), not enough evidence
  (at the 95% confidence level) to reject the NULL Hypothesis.

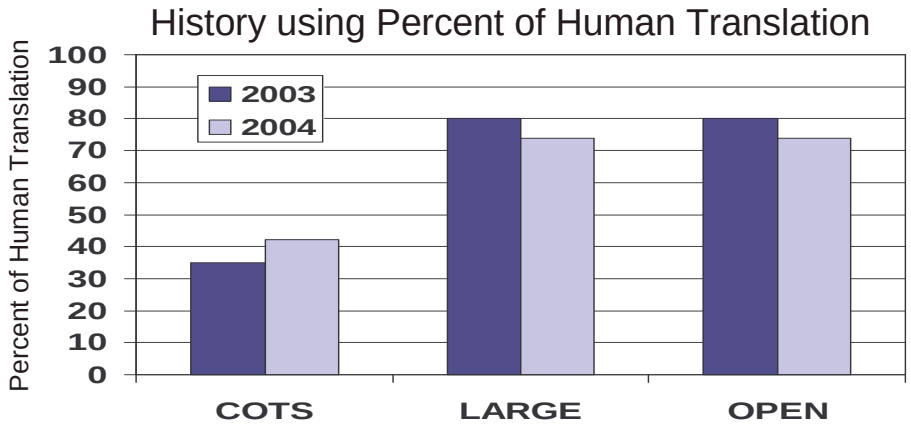
Sign Test Results: Chinese
Primary Systems for the LARGE data track

System Y	System X									
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Y1 (1)		>99.9%	<0.1%	2%	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%
Y2 (0)			<0.1%	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%
Y3 (6)				>99.9%	>99.9%	<0.1%	>99.9%	<0.1%	<0.1%	99.8%
Y4 (2)					<0.1%	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%
Y5 (3)						<0.1%		<0.1%	<0.1%	<0.1%
Y6 (9)							>99.9%	>99.9%	>99.9 %	>99.9%
Y7 (3)								<0.1%	<0.1%	<0.1%
Y8 (7)										>99.9%
Y9 (7)										>99.9%
Y10 (5)										

- Answers the question:
  What is the confidence that system Y is better than system X?
- If no entry (yellow cells), not enough evidence
  (at the 95% confidence level) to reject the NULL Hypothesis.

History – Arabic (Newswire)

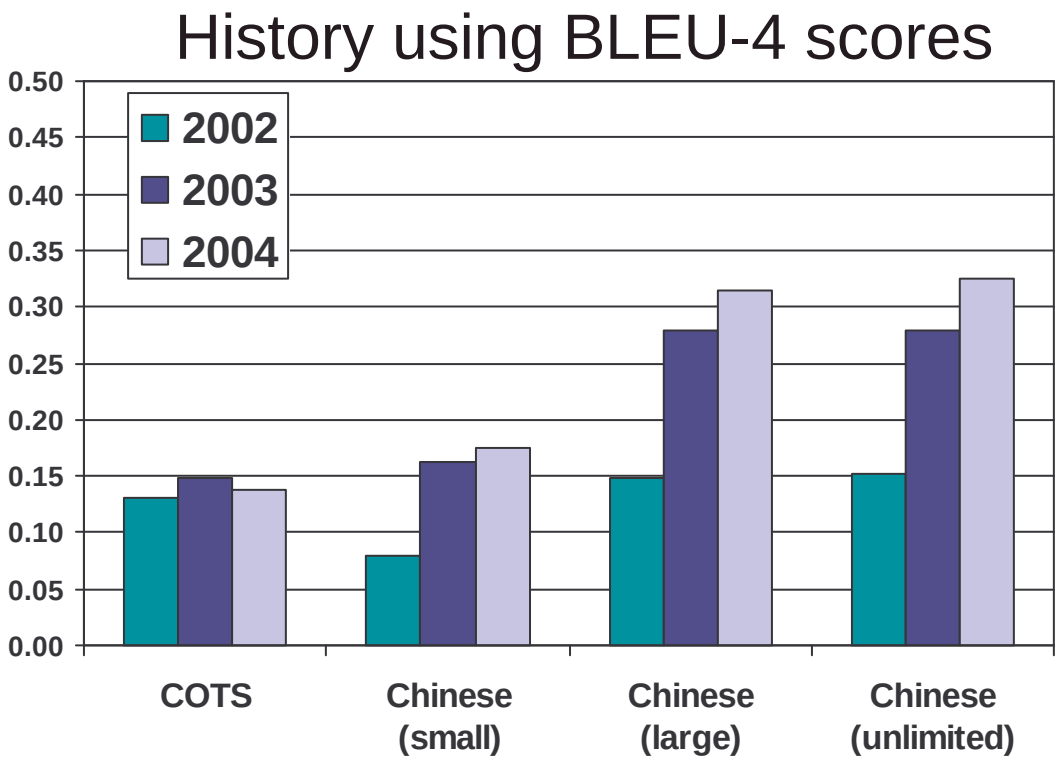


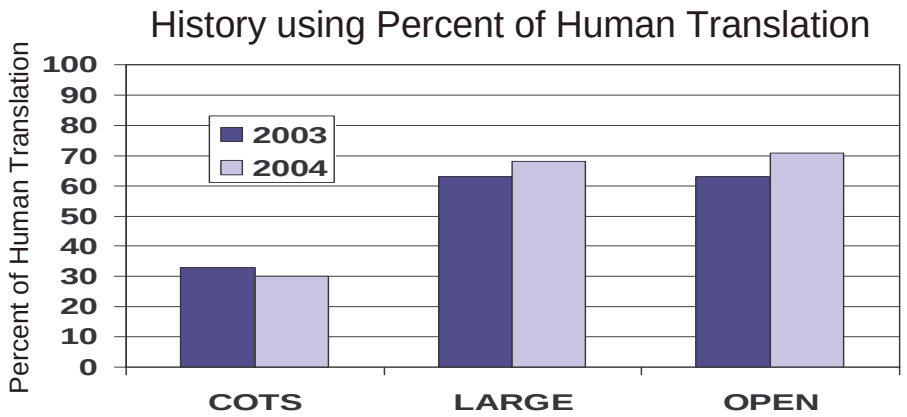


- Best on-time "evaluation" system (primary or contrastive)
- Scored using
  - mteval-v11.pl
  - Case sensitive
- COTS is web-base (may be updated)

- Improvement for overall BLEU scores, but ***not*** for "*Percent of Human Translation" measure*

History – Chinese (Newswire)





- Best on-time "evaluation" system (primary or contrastive)
- Scored using
  - mteval-v11.pl
  - Case sensitive
- COTS is web-base (may be updated)

- Improvement for overall BLEU scores ***and*** the "Percent of Human Translation" measure

Summary

- Record level of participation
  - Contrastive site results are listed in the release
- Official metric is BLEU (4-gram)
- MT-04's test data included two new genres
  - Editorials
    - Arabic – difficult
    - Chinese – comparable to newswire
  - Speeches
    - Arabic – more difficult than newswire
    - Chinese – least difficult genre tested
- Improvement noted over last year when limiting to the common genre – newswire
- Evidence to suggest more similar reference translations for Arabic in 2004
  - (possibly influencing the percent of human scores?)